Invited Forum: Challenges in Making Data Available
General Article

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(1) 121–130

DOI: 10.1177/2515245917746500
www.psychologicalscience.org/AMPPS

Practical Solutions for Sharing Data and Materials From Psychological Research

Rick O. Gilmore[1], Joy Lorenzo Kennedy[2], and Karen E. Adolph[3]
[1]Department of Psychology, The Pennsylvania State University; [2]Databrary.org, New York University; and [3]Department of Psychology, New York University

Abstract
Widespread sharing of data and materials (including displays and text- and video-based descriptions of experimental procedures) will improve the reproducibility of psychological science and accelerate the pace of discovery. In this article, we discuss some of the challenges to open sharing and offer practical solutions for researchers who wish to share more of the products—and process—of their research. Many of these solutions were devised by the Databrary .org data library for storing and sharing video, audio, and other forms of sensitive or personally identifiable data. We also discuss ways in which researchers can make shared data and materials easier for others to find and reuse. Widely adopted, these solutions and practices will increase transparency and speed progress in psychological science.

Keywords
data sharing, ethics, video and audio recordings

Received 9/7/17; Revision accepted 11/14/17

Psychological science faces a crisis: Many findings cannot be readily reproduced, and progress is slow (Open Science Collaboration, 2015; Shrout & Rodgers, in press). To improve reproducibility and accelerate discovery, researchers are calling on their colleagues to share their raw data and materials (Munafò et al., 2017). Greater openness allows for reproduction of original findings, identification of contextual variables that can lead to different patterns of results, and secondary reuse of the data to answer new questions. In many cases involving research about human behavior, personally identifying information can be easily removed, and sharing poses no risk to participants. But sharing sensitive data or data that contain identifiers that cannot be easily removed poses real challenges in many areas of psychological science.

In this article, we describe practical solutions for researchers who are considering whether to share their data, materials, procedures, and analyses. Our recommendations stem from the solutions we devised in building Databrary (databrary.org), a Web-based data library specialized for storing and sharing video with the research community. Databrary's approach to data stewardship overcomes hurdles to sharing personally identifying information—for example, the faces, voices, and home or classroom interiors recorded by video—by building on established practices (e.g., Bloomrosen & Detmer, 2008) and foundational ethical principles such as informed consent. In addition, we discuss how sharing data and materials in standardized, searchable ways maximizes the potential for future reuse.

Curating and Sharing Research Products

Sharing research products beyond the figures and statistical summaries found in published articles enables researchers to evaluate the strength of existing claims, replicate published findings, build on others' expertise

Corresponding Author:
Rick O. Gilmore, Department of Psychology, Moore Building, The Pennsylvania State University, University Park, PA 16802
E-mail: rogilmore@psu.edu

and experience, and reuse data to make new discoveries. Indeed, shared data sets (e.g., AddHealth, CHILDES/TalkBank; the Early Childhood Longitudinal Study) have spawned large and productive research communities (Gilmore, 2016). Researchers who study children's language have long traditions of open sharing (MacWhinney, 2001), but other communities do not. Researchers considering whether to share data, materials, and analyses face several questions: What should I share? Where and with whom should I share? When should I share? And how do I share? In this section, we provide some answers.

What to share

A growing consensus among researchers is that authors should share at least the data files underlying the statistical findings described in their publications. Indeed, increasing numbers of journals require this, and many research funders (e.g., the U.S. National Science Foundation and the Gates Foundation) expect it. Data with personal identifiers can be shared by removing or altering these elements or by securing participants' explicit permission to share the data.

In addition, researchers can share materials that enrich the meaning and utility of research data: behavioral tasks and test instruments (if they are unencumbered by intellectual-property restrictions), detailed empirical protocols and code books, computer code (e.g., SPSS or SAS syntax; MATLAB, R, or Python scripts), and images from brain-imaging studies. Of course, some components cannot be shared in their entirety or in an unaltered state. Such materials include proprietary data or software that is restricted by legal or contractual obligations, sensitive data that could cause participants harm, and personally identifiable data for which permission to share has not been granted.

Where and with whom to share

Researchers face many choices in deciding where to share data and materials. Options include personal, lab, and project-specific Web sites; institutional repositories; open-science services (e.g., the Open Science Framework, or OSF); data repositories (e.g., Databrary; Dryad; Dataverse; the Inter-university Consortium for Political and Social Research, or ICPSR; TalkBank; and WordBank), and supplemental materials attached to an article and stored on a publisher's Web site (see Table 1).

The decision about where to share is linked to the question of who can and should have access to the data. In deciding among storage options, researchers should ask themselves how accessible they want the data or materials to be—publicly accessible, accessible to a community of researchers, or accessible only to researchers who are specially selected or vetted. Shared data need not be made publicly available to meet open-science standards or journal or funder mandates. In fact, some of the most successful examples of data sharing in the social and behavioral sciences involve *restricted access*, in which data are stored in recognized data repositories that limit access to researchers. Public access may be appropriate for some data sets, but it should not be the standard to which all studies must be held.

For many psychological scientists, storing data and materials alongside a published article, or in a repository that is linked to a published article, will maximize the visibility and discoverability of those data and materials and their potential for reuse. Repositories have distinct advantages. Most are operated by not-for-profit research institutions that have open information sharing as a core mission. For example, ICSPR is hosted at the University of Michigan, TalkBank is hosted at Carnegie Mellon University, Databrary is hosted at New York University, and WordBank and OpenNeuro are hosted at Stanford University. Storage in a repository increases the likelihood that shared data and materials can be easily found and repurposed by others, that they will be preserved for the long term, and that sharing will garner citations by researchers who reuse the data or materials. Moreover, sharing information in a repository relieves the researcher from the obligation of having to address individual requests for sharing on a case-by-case basis or sorting through how to transfer files when changing institutions.

A given set of data and materials can be stored in multiple places, of course, but this duplicates work and increases the burden on researchers. One solution is to choose a central hub for a study (e.g., OSF, ICPSR, Dataverse, or Databrary) and provide links to other repositories or Web services with features appropriate for specialized types of data or materials (e.g., Bergelson, 2017; Gilmore, 2014). For example, Databrary encourages researchers to store videos and associated materials on the Databrary site and include links to neuroimaging data shared elsewhere. Other repositories and Web-based analysis tools specialize in storing, providing visualizations of, and analyzing neuroimaging data (Gilmore, Diaz, Wyble, & Yarkoni, 2017; Gorgolewski et al., 2016; Poldrack & Gorgolewski, 2017; Poldrack et al., 2017; Yarkoni, Poldrack, Nichols, Essen, & Wager, 2011). Similarly, there are excellent sites (e.g., GitHub; see Table 1) for hosting data-analysis code and reproducible version-controlled analysis workflows (e.g., Seisler & Gilmore, 2017).

Table 1. Some Resources for Sharing Data and Materials

Resource	Comments
Databrary (http://databrary.org)	Public sharing or restricted sharing with institutionally authorized researchers; video and audio recordings, documents, coding files
Dataverse (http://dataverse.org)	Public or restricted sharing of many types of data and materials
Dryad (http://datadryad.org/)	Public sharing of data sets and scripts associated with specific publications
figshare (https://figshare.com)	Public sharing of graphs, figures, and oral presentations
GitHub (https://github.com)	Public sharing of research materials, data, and code
Inter-university Consortium for Political and Social Research (https://www.icpsr.umich.edu/icpsrweb/)	Restricted and unrestricted sharing of multiple types of data
National Database for Autism Research (http://ndar.nih.gov)	Largely unrestricted sharing of a wide range of behavioral and biological data from studies focusing on autism spectrum disorder
OpenNeuro (http://openneuro.org)	Public sharing of brain-imaging data sets
Open Science Framework (http://osf.io)	Public sharing of multiple types of data and research materials; preregistration of research plans
Protocols.io (http://www.protocols.io)	Open (public) or private sharing of research protocols
TalkBank (http://talkbank.org)	Open sharing of audio and video recordings of language samples and speech transcripts; includes population- and measure-specific collections (e.g., HomeBank, CHILDES)
WordBank (http://wordbank.stanford.edu)	Public sharing of MacArthur-Bates Communicative Development Inventory (Fenson et al., 2007) data and metadata
Zenodo (https://zenodo.org)	Sharing of research outputs within a self-curated community

When to share

Despite a lack of consensus about when in the workflow data and materials should be shared, open-science advocates and research funders say "never" is simply too late. Several standards have emerged: Many journals advocate sharing when an article goes to press, but other recommendations are that sharing should take place within 3 years after deposit into a repository (e.g., OpenNeuro) or at the end of a grant period (e.g., The Human Connectome Project). Both the OSF and Databrary allow researchers to upload data and materials at any point in the research and publication process and keep them private until the team is ready to share. If our field can arrive at a consensus about a reasonable timeline for sharing, including a consensus for sharing longitudinal data sets from multiple waves of collection, we may prevent the imposition of standards from the outside.

How to share

Best practices for sharing data are known as the FAIR principles (Wilkinson et al., 2016): That is, data should be shared in *f*indable, *a*ccessible, *i*nteroperable, and *r*eusable forms. Similar principles apply to sharing materials. Findability and accessibility primarily have to do with where items are stored. Both are facilitated by repository catalogues such as the one maintained by the Data Preservation Alliance for the Social Sciences, also known as Data-PASS (http://www.data-pass.org). Interoperability and reusability, on the other hand, concern file formats. Some fields have established data standards (e.g., CHAT for language transcription) or are creating them (e.g., the Brain Imaging Data Structure, or BIDS, in neuroscience). In other fields, no standards exist; in these cases, following FAIR principles means sharing text files, rather than spreadsheets, PDFs, or MS Word documents. Because so much of a data analyst's time is spent cleaning data, a growing number of data scientists advocate storing data in "tidy" or "long" formats (Wickham, 2014) accompanied with full data dictionaries. Finally, the data most valuable for reuse are those shared in as raw a form as possible. This means sharing data at the participant, or even event or trial, level, rather than sharing only group summaries. In the case of audio and video data, this means sharing the actual recordings, not just processed coding files.

Planning for sharing

Answering these "wh" questions about sharing research data, materials, and procedures may seem daunting to researchers who are new to these practices. But in reality, either implicitly or explicitly, most researchers already confront similar questions when planning a study and seeking approval of an institutional review board (IRB) or other ethics board. We advocate that

researchers make "planning for sharing" an explicit part of their ongoing research process and that this should include planning for sharing data, materials, and procedures.

Meeting the Challenges of Sharing Identifiable Information

We now turn to the challenges associated with sharing identifiable information, especially video recordings, and how Databrary has resolved these issues. We note that Databrary's specific solutions to these problems have general relevance for research that does not involve video or audio recordings or the collection of sensitive data. Researchers who use video recordings face particularly difficult challenges in balancing adherence to the principles of ethical research with the desire to share openly. Video provides an incomparably rich source of information about human behavior and an unrivaled means of documenting empirical procedures (Adolph, 2016; Adolph, Gilmore, & Kennedy, 2017; Gilmore & Adolph, 2017; Suls, 2013). However, sharing and reusing video data pose ethical challenges (e.g., how to protect participants' privacy), technical challenges (e.g., how to store large video files), practical challenges (e.g., how to find relevant files), and scientific challenges (e.g., how to annotate video). The ethical challenges prove the thorniest because video inherently contains personally identifying information—participants' faces and voices, their names spoken aloud, and views of their homes or classrooms. Although altering recordings to protect participants' identities is possible, blurring or obscuring faces and voices sharply diminishes their value for reuse. For example, the analysis of emotion is hugely compromised without clear views of the face, as are analyses of linguistic inputs based on altered or redacted audio.

The question of how to protect sensitive or identifiable data is not new. When the Databrary project began in late 2012, we built upon a range of established best practices at other repositories in the social and behavioral sciences. Databrary drew particular inspiration from TalkBank at Carnegie Mellon University, which has been hosting video and audio recordings of human speakers for many years (MacWhinney, 2001), and from ICPSR, which hosts classroom videos from the Methods of Effective Teaching project. Our aim, and that of our sponsors (Eunice Kennedy Shriver National Institute of Child Health and Human Development and the National Science Foundation), was to create a system that built on the best ideas from TalkBank and ICPSR but allowed the storage and sharing of video data from a broad range of human behaviors and contexts (e.g., labs, homes, classrooms, museums) and across many diverse individual studies. We focused on the developmental and learning sciences, in which video recordings are commonplace. We sought to devise a policy framework that allowed unaltered videos to be shared as widely and openly as possible, while minimizing the risk of privacy violations. To achieve this, we consulted extensively with TalkBank, ICPSR, experts on Databrary's Advisory Board, and officials at New York University and the Pennsylvania State University who had expertise in legal issues, library science, privacy and cyber security, research ethics, and sponsored-projects management. The two-pronged framework that emerged (a) restricts access to identifiable data to researchers who have explicit authorization and ethics oversight by their institutions and (b) requires contributors to obtain participants' (or their parents') permission to share identifiable data. Table 2 provides information about resources for researchers who wish to obtain permission to share identifiable data and recordings.

Table 2. Resources for Seeking Permission to Share Identifiable Data and Recordings

Source	Resource
Databrary	Institutional access agreement (https://www.databrary.org/access/policies/agreement.html)
Databrary	Template for obtaining permission to share (https://www.databrary.org/resources/templates/release-template.html)
Inter-university Consortium for Political and Social Research	Recommendations for informed-consent language (https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html)
Open Brain Consent	Consent document for neuroimaging research (https://open-brain-consent.readthedocs.io)
Open Humans Project	Informed-consent document (https://www.openhumans.org/static/public-data/docs/Consent_Document_20160128_(stamped).005ab78912c1.pdf)
TalkBank	Information for seeking permission to contribute data to TalkBank from an institutional review board or other ethics board (http://talkbank.org/share/irb/)
U.S. government	New Common Rule language (https://www.federalregister.gov/documents/2017/01/19/2017-01058/federal-policy-for-the-protection-of-human-subjects)

Restricted access via institutional agreement

Institutions regularly execute data-use agreements to govern sharing of sensitive or valuable data, but such agreements are uncommon in some areas of psychological science. Most data-use agreements give a specific user access to a particular data set for a limited purpose. We wanted to create a "data commons" to make data sharing more open, less restrictive, and more scientifically generative while maintaining the important legal protections that conventional data-use agreements provide. Based on models shared by ICPSR, the Databrary Access Agreement (https://www.databrary.org/access/policies/agreement.html) formalizes the relationship between researchers, their institutions, and Databrary's host institution, New York University. The agreement allows researchers to upload and store data and materials on Databrary, and to use Databrary for sharing data and materials that they have permission to share (from their participants, their institutions, and their governments, if applicable). The agreement also grants researchers access to all shared data and materials in Databrary for any research use that has approval of an IRB or other ethics board, as well as for preresearch, educational, and noncommercial uses. Thus, the Databrary model is broader than typical data-use agreements. It authorizes researchers to contribute information themselves, to use data and materials contributed by others, and to use shared information for multiple purposes.

As a binding legal document between institutions, the access agreement must be executed by an official with signing authority (typically, an official in the university's grants or contracts office). Although some researchers may be surprised to learn that their institutions "own" the products of the research they worked so hard to write grants for, including the data they collect, analyze, and describe, institutional ownership is the reason why Databrary's agreement requires institutional approval (as do comparable data-use or material-transfer agreements). In signing the agreement, institutions attest that they maintain an ethics review board or, in some international cases, that they require researchers to submit research proposals to an external ethics board. In this way, the agreement addresses intellectual-property and research-ethics concerns.

Although the Databrary Access Agreement legally binds an institution, researchers also sign it. In cosigning the agreement, researchers promise to (a) respect participants' wishes about sharing data, (b) treat other researchers' data with the same high standards of care that they use with their own research data, and (c) take responsibility for the ethical behavior of other people (students, staff, colleagues) to whom they grant Databrary access. Institutionally approved researchers, called *Authorized Investigators*, will often want to grant some level of access to other people, called *Affiliates*, and Authorized Investigators must take full responsibility for those Affiliates. All Authorized Investigators seeking access to Databrary must complete training in human-participants research ethics; Affiliates must also complete ethics training if required by their sponsoring Authorized Investigator or their institution. Databrary's language regarding the required ethics training is purposely broad to accommodate specific institutional requirements and local ethical norms. Approval of an IRB or other ethics board is required both to share information on Databrary and to reuse the information stored there for research purposes. Databrary's approval is not required to browse shared research videos, watch videos about research procedures (e.g., how to recruit and test participants, obtain informed consent, and use software or specialized equipment), engage in other preresearch activities, or use shared videos for teaching or in research presentations. Table 3 provides information about how to access, share, and reuse data and materials shared on Databrary.

Executing the Databrary Access Agreement may seem like an extra and possibly burdensome step. Indeed, OSF, Dataverse, Dryad, and OpenNeuro have no parallel requirement. However, none of these services are designed for storing or sharing identifiable or sensitive data. The Databrary agreement provides increased protection to participants, researchers, and their institutions. It provides a practical solution for researchers who collect sensitive or identifiable data and want to share it, and for those who want to reuse sensitive or identifiable data shared by others. The agreement also speaks to concerns about the possible risks of sharing "de-identified" data without restrictions on who can access it, particularly given doubts about whether procedures for removing identifiers from data truly protect participants' identities (Cavoukian & Castro, 2014; Narayanan, Huey, & Felten, 2016; Ohm, 2009).

The Databrary agreement unites a growing international network of institutions (currently 367 strong), Authorized Investigators (currently 666), and Affiliates (currently 293) in a research community bound to a common set of principles. The agreement embraces the virtues (and legal protections) of data-use agreements, but extends the scope beyond single studies and is more uniform and less restrictive. Authorized Investigators can freely access, use, and reuse shared data and materials in Databrary as long as participants' permission levels are respected and the original sources are properly cited. In this way, Databrary strikes a balance between providing open access to research data and the ethical imperative to protect participants' privacy.

Table 3. Steps for Accessing, Sharing, and Reusing Data and Materials on Databrary

Use and step	Description
Preresearch or educational use	
Register	All users must create an account at http://databrary.org/register.
Secure authorization	PIs request authorization from an institution; students, lab staff, and postdoctoral researchers can request authorization from an authorized PI.
Browse data or materials	Following authorization, users can stream or download data for nonresearch, preresearch, or educational uses; approval of an IRB or other ethics board is not required at most institutions.
Sharing self-collected data	
Seek IRB or ethics-board approval	The research team must obtain permission from an IRB or other ethics board to collect personally identifiable video data and share it with Databrary.
Seek participants' permission to share	Using Databrary's release template or equivalent language, the team must document each participant's level of permission to share data (see Table 4).
Upload data to Databrary	Videos and project-, session-, and participant-level metadata can be uploaded while a study is in progress or after completion of the study. Links to files at external resources (Open Science Framework, GitHub, OpenNeuro, etc.) can be added.
Share data	When the research team chooses (e.g., a report is published or at the end of the grant period), the team can share the volume, granting other researchers (other Databrary investigators or the public) access to the files.
Accessing shared data to conduct research	
Register	All users must create an account at http://databrary.org/register.
Secure authorization	PIs request authorization from an institution; students, lab staff, and postdoctoral researchers can request authorization from an authorized PI.
Seek IRB or ethics-board approval	The research team must obtain permission from an IRB or other ethics board to reuse personally identifiable data.
Download data	The team can search for and filter shared data according to the task or participant characteristics of interest and then download the data for analysis.
Share the reanalyzed data	The team can upload any new data and add links to external resources (Open Science Framework, GitHub, OpenNeuro, etc.) and Databrary data sets that were used.

Note: IRB = institutional review board; PI = principal investigator.

From informed consent to permission to share

Informed consent has been a central tenet of research ethics involving human participants for decades. The second prong of Databrary's policy framework extends this principle to include data sharing. The risks of sharing data often differ from those associated with participation in research. Participation may involve physical or psychological risks that data sharing does not because data sharing does not involve direct contact with participants. However, sharing videos or other identifiable information poses risks to participants' privacy and the confidentiality of the data they provide. Someone not supervised by the original research team might identify a participant and reveal his or her participation, along with other information, some of which may be sensitive. Indeed, one of the motivations for the U.S. government's new Common Rule regulations (Federal Policy for the Protection of Human Subjects, 2017), effective January 19, 2018, was the increasing use of digital records, such as video records, and the need to add new rules governing "informational" harm to the existing rules governing physical risk.

As discussed in the previous section, Databrary mitigates this risk by restricting access to institutionally authorized researchers who have ethics training and the affiliates they supervise and assume responsibility for. In addition, Databrary requires that researchers obtain participants' explicit permission to share videos and other identifiable data and indicate the level of permission on Databrary. This ensures that participants make informed choices about the potential risks of sharing. Although the Databrary model is new, depending on the individual IRB or ethics board and the terminology in the original research consent form, it is sometimes possible to grandfather into Databrary recordings collected long before Databrary's consent model was created (e.g., Arnold Gesell's archival films from the 1930s and 1940s; Baker, 2014).

Typically, obtaining participants' permission to share data differs from obtaining their consent to participate in a research study. However, IRBs and other ethics boards differ in the approach they prefer, so three general models for securing permission to share data have emerged:

- Model 1: consent to participate in research and permission to share data are considered completely separate choices and are recorded on separate documents that are part of one research protocol
- Model 2: permission to share data is the focus of a completely separate research protocol that covers multiple projects; a participant choosing to participate in research and share data must consent to two protocols
- Model 3: consent to participate in research and permission to share data are integrated into one all-inclusive consent document

Each model has virtues and flaws. The downside of combining research consent and permission to share, as in Model 3, is that participants can be lost to a study if they refuse to share their data. In other words, requests to share data can impede recruitment. Thus, Databrary recommends Model 1 or Model 2, to keep the decisions separate. Although consent to participate must obviously be given prior to participation, a best practice for permission to share data is to seek it *after* the completion of research activities. This ensures that participants are fully aware of the study's procedures and what they are being asked to share.

In practice, we find that participants' willingness to share data exceeds researchers' expectations, regardless of how that permission is sought. The vast majority of participants willingly—and often eagerly—agree to share video and related identifiable data. Many types of behavioral research are not particularly sensitive from the participants' point of view, and this is especially true of studies that involve recording videos similar to those people share on social media. Moreover, most volunteer research participants want their participation to generate the maximum benefit to science, and they understand that sharing their data helps to realize that ideal. Parents of children with disabilities are especially eager to share data in the hope that it can speed progress toward successful interventions and increased understanding.

The Virtues of Standardization

One of the most important accelerators for scientific progress is the adoption of standard practices and metrics. Researchers must speak a common language to ensure that participants are protected and to minimize the burdens of data curation and sharing. Databrary has standardized several aspects of the sharing process.

Standardizing permission to share

Databrary has standardized the way permission to share is sought and recorded (Table 4) to create consistent rules regarding who can access shared data. Release documents commonly used by researchers for permission to share photos and video recordings served as the starting point in developing these policies. Those releases, often in checklist form, document permission to use photos or video clips from raw research videos during oral presentations, in publications, on lab Web sites, or in textbooks. However, release language is idiosyncratic to individual researchers. This creates problems for sharing because different uses pose very different risks of disclosure. Posting a video clip on an unrestricted Web site, so that the file can be downloaded and redistributed by anyone, poses a greater disclosure risk than storing a clip in a restricted repository (where the original file remains under the control of a researcher) but allowing authorized users to show the clip in class. For open sharing of video data to succeed, everyone—researchers and participants alike—must have the same

Table 4. Databrary's Release Levels for Data Sharing

Level	Explanation
Unreleased	Signed data-sharing releases were not obtained or are unavailable; the data are available only to the research team.
Private	The participant said "no" to data sharing or the researcher has chosen not to share the data; they are available only to the research team.
Authorized users	The data are available on Databrary to authorized researchers and their affiliates.
Excerpts	The data are available on Databrary to authorized researchers and their affiliates; clips may be shown during oral presentations or for teaching purposes.
Public	Anyone may stream or download the data.

understanding about what is, and what is not, allowed. Therefore, Databrary created a set of data-sharing release levels that map onto the different levels of disclosure risk that are often obscured by video or photo release checklists. Databrary's release levels are unique, but we advocate the widespread adoption of these or comparable standards.

At the most conservative end of the sharing spectrum, files are marked *unreleased* to indicate that signed releases were not obtained or are otherwise unavailable. Note that this label does not indicate that the participant said "no" to sharing; rather, it indicates that the participant's wishes about sharing are unknown. An unreleased video is available only to the researchers who contributed it and to any specific collaborators they choose. The recordings are stored on Databrary to keep data sets complete, to facilitate collaboration, and for long-term archival purposes.

Data are marked *private* when the participant said "no" to sharing or the researcher has chosen not to share. These files are also available only to the original research team and any collaborators. Private and unreleased files are marked with general demographic information so that other researchers can determine whether the shared data are representative.

Data are marked for *authorized users* when participants gave permission to share their data with other researchers, including Authorized Investigators and any Affiliates they grant access to. Ideally, all videos on Databrary would be shared at least at this level, to facilitate reproducibility and reuse among the Databrary community.

Data are marked for *excerpts* when participants gave permission to share their data with authorized users, and those users may show portions of the data (short video clips or photos) in presentations (e.g., classroom lectures, research conferences, colloquia) for instructional or informational purposes. Although excerpts are available for download only to authorized Databrary users, presentations using the clips could be videotaped or recorded, and those recordings might then be released into the public domain (e.g., a YouTube video of a conference presentation). The essential point is that an authorized person chooses the time, place, and manner for showing a clip to what may be an audience more public than a lab group.

Finally, data are marked *public* when participants placed no restrictions on the sharing of their data and agreed that full videos or clips can be used by anyone for any purpose.

Implementing standard release levels

Researchers control the assignment of release levels to their data and materials. They may choose to share data at a more restrictive level than participants granted, but may never choose a less restrictive level. The five release levels are file-level settings; each video or data file must be marked with a release level, and *private* is the default.

Files are stored on Databrary in volumes (coherent collections of data, with any supporting materials the researchers wish to include) that are by default unshared, visible only to Authorized Investigators or Affiliates specifically chosen by the contributing researchers. Unshared volumes will not appear in a search, even when it is conducted by an authorized Databrary user. A volume can contain files with a mixture of release levels according to the permission granted by each participant. So, once a volume is shared, files marked *unreleased* or *private* continue to be viewable or downloadable only by people selected by the research team. This feature can be particularly useful for publicly sharing coding manuals, procedural videos, and other research materials that contain no personally identifiable information and therefore can be shared at a less restrictive level than files with participants' data. Note that an unshared volume can contain files marked for eventual sharing at the authorized-users and excerpts levels. This typically occurs when some of the collaborating researchers on a project are not yet ready to share the volume.

Databrary makes available exemplar videos and scripts for obtaining participants' permission to share their data (https://www.databrary.org/resources/guide/investigators/release/asking/examples.html, https://www.databrary.org/resources/templates/release-template.html). Using these templates helps to increase the consistency of the language used and the procedure that is followed.

Standardizing demographic metadata

In addition to standardizing permission language and release levels, Databrary has sought to standardize the study-, task-, session-, and participant-level metadata that accompany a data set. We recognized that the formats psychological scientists use to code dates, demographic information, and vital task-related information lack consensus standards, so we created a spreadsheet-like interface to encourage more systematic and standardized recording of essential data about participants' characteristics (e.g., age at test, birth date, gender, race-ethnicity), the test session (e.g., geographic location, location type), measures, and study conditions. Entering these data facilitates consistency across collaborators at different research sites. In addition, by entering these data and metadata into Databrary, researchers make it easier for others to search across the library, find data that suit their specific needs, combine data with similar characteristics, and reuse shared information.

We think that all researchers embracing open-science practices should share essential metadata about their participants, testing sessions, measures, and methods in standardized ways that will allow shared data sets to be combined into robust structures of knowledge. Although some researchers may resist standardization of data formats, standardization has compelling virtues. The CHAT data format developed for use in CHILDES and the TalkBank family of repositories (MacWhinney, 2001) created a universal "language" for meaningfully tagging human speech that provides a foundation for analyses within and across data sets that could not proceed otherwise. More recently, we find especially encouraging the progress being made by researchers who are forging the BIDS standard in neuroimaging (Gorgolewski, et al., 2016). Neuroimaging data sets stored in the BIDS format can be readily imported into standardized Web-based repositories. Further, the use of a consistent file structure enables scriptable, reproducible brain-imaging analysis pipelines. We also draw inspiration from WordBank (Frank, Braginsky, Yurovsky, & Marchman, 2017). This repository allows researchers to use a browser to visualize and manipulate data from a standard parent-report measure of children's vocabulary development, the MacArthur-Bates Communicative Development Inventory (M-CDI; Fenson et al., 2007). Some WordBank M-CDI data are accompanied by metadata about the child, parents, and family, and when these metadata are available, the system can be used for powerful exploratory analysis. TalkBank, Open Neuro, and WordBank demonstrate that investments in creating standard ways of encoding behavior, participants' characteristics, and study metadata can yield concrete payoffs.

Conclusion

The new Common Rule states that "the scientific community recognizes the value of data sharing and open-source resources and understands that pooling intellectual resources and capitalizing on efficient uses of data and technology represent the best ways to advance knowledge" (Federal Policy for the Protection of Human Subjects, 2017, § 1A). In that spirit, open-science advocates should minimize the cost of data sharing to individual researchers while maximizing benefits to the field as a whole and to the public. Advance planning for sharing—determining what data will be shared, how, where, and in what form—can help reduce costs to investigators, or at least spread costs across the entire trajectory of a study, while maximizing the visibility and reuse potential for these products of scholarship. Sharing data and materials in standardized, searchable forms in repositories can serve multiple purposes.

Databrary is one of a family of data- and materials-sharing tools available for psychological scientists seeking to share more information more widely. Databrary demonstrates that it is possible to openly share unaltered research videos with personally identifiable information while still protecting participants' privacy. Moreover, Databrary facilitates research transparency by making readily viewable the subtle details of procedures and materials (Adolph et al., 2017; Gilmore & Adolph, 2017), a practice that other large-scale replication efforts (e.g., ManyLabs 4, n.d.; ManyBabies, Frank et al., in press) have begun to adopt. Databrary also accelerates progress by allowing researchers to exploit the richness of video to answer new questions using data already collected by others (Adolph, 2016). The rapid growth of the Databrary community over the past several years demonstrates the feasibility of this framework for sharing video, and we argue that it can be extended to research programs that collect other personally identifiable or sensitive information.

The history of large-scale data sharing in the developmental sciences (Gilmore, 2016) demonstrates that capturing data and metadata about participants, testing sessions, tasks, and measures in consistent ways makes shared data maximally valuable for secondary reuse by other researchers. Beyond serving the laudable goal of increasing transparency, data sharing and expanded data reuse will accelerate discovery. Of course, the greater the amount of data that is stored, the greater the possible risk to participants. And that reinforces the need to adopt practices that encourage data sharing, but with specific restrictions on access and consistent requirements for securing participants' permission.

We hope that the Databrary model provides useful ideas for other researchers interested in adopting open-science practices and a home for those researchers who want to store and share video files as raw data or procedural documentation (Adolph, 2016; Adolph et al., 2017; Gilmore & Adolph, 2017). We look forward to continued conversations about how to implement practical solutions to the challenges of sharing research data and materials in ways that advance discovery in psychological science.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

J. L. Kennedy created the original draft of the manuscript. R. O. Gilmore and K. E. Adolph made substantial revisions. R. O. Gilmore prepared the final document for submission.

Declaration of Conflicting Interests

R. O. Gilmore and K. E. Adolph are the cofounders and co-directors of Databrary. J. L. Kennedy is Databrary's Scientific

Support Specialist. The authors declared that there were no other potential conflicts of interest with respect to the authorship or the publication of this article.

Funding

The authors acknowledge support from the National Science Foundation (BCS-1238599), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (U01-HD-076595), the Society for Research in Child Development, and the Alfred P. Sloan Foundation.

References

Adolph, K. E. (2016). Video as data: From transient behavior to tangible recording. *Observer*, *29*(3), 23–25.

Adolph, K. E., Gilmore, R. O., & Kennedy, J. L. (2017, October). Video data and documentation will improve psychological science. *Psychological Science Agenda*. Retrieved from http://www.apa.org/science/about/psa/2017/10/video-data.aspx

Baker, D. (2014). *Arnold Gesell's films of infant and child development* [Data set]. Retrieved from http://doi.org/10.17910/B7.70

Bergelson, E. (2017). *SEEDLingS 6 Month* [Data set]. Retrieved from http://doi.org/10.17910/B7.330

Bloomrosen, M., & Detmer, D. (2008). Advancing the framework: Use of health data—a report of a working conference of the American Medical Informatics Association. *Journal of the American Medical Informatics Association*, *15*, 715–722. doi:10.1197/jamia.M2905

Cavoukian, A., & Castro, D. (2014). *Big data and innovation, setting the record straight: De-identification does work*. Toronto, Ontario, Canada: Information and Privacy Commissioner.

Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. (2017). Retrieved from https://www.federalregister.gov/documents/2017/01/19/2017-01058/federal-policy-for-the-protection-of-human-subjects

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *MacArthur-Bates Communicative Development Inventories*. Baltimore, MD: Paul H. Brookes.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., . . . Yurovsky, D. (in press). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). WordBank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*, 677–694. doi:10.1017/S0305000916000209

Gilmore, R. O. (2014). *Children's brain responses to optic flow vary by pattern type and motion speed* [Data set]. Retrieved from http://doi.org/10.17910/B7QG6W

Gilmore, R. O. (2016). From big data to deep insight in developmental science. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*, 112–126.

Gilmore, R. O., & Adolph, K. E. (2017). Video can make behavioural science more reproducible. *Nature Human Behavior*, *1*(7), Article 0128. doi:10.1038/s41562-017-0128

Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1396*, 5–18. doi:10.1111/nyas.13325

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., . . . Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data, 3*, Article 160044. doi:10.1038/sdata.2016.44

MacWhinney, B. (2001). From CHILDES to TalkBank. In B. MacWhinney, M. Almgren, A. Barreña, M. Ezeizaberrena, & I. Idiazabal (Eds.), *Research in child language acquisition* (pp. 17–34). Somerville, MA: Cascadilla Press.

Many Labs 4: Investigating effects of researcher expertise on replication outcomes. (n.d.). Retrieved from https://osf.io/8ccnw/

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. doi:10.1038/s41562-016-0021

Narayanan, A., Huey, J., & Felten, E. W. (2016). A precautionary approach to big data privacy. In S. Gutwirth, R. Leenes, & P. De Hert (Eds.), *Data protection on the move* (pp. 357–385). Dordrecht, The Netherlands: Springer.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701–1777.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., . . . Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*, 115–126. doi:10.1038/nrn.2016.167

Poldrack, R. A., & Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *NeuroImage*, *144*(Part B), 259–261. doi:10.1016.neuroimage.2015.05.073

Seisler, A. R., & Gilmore, R. O. (2017). Developing R code for the processing and analysis of optic flow data. In J. Kitzes, D. Turek, & F. Deniz (Eds.), *The practice of reproducible research: Case studies and lessons from the data-intensive sciences* [Online version]. Retrieved from http://www.practicereproducibleresearch.org

Shrout, P. E., & Rodgers, J. L. (in press). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*.

Suls, J. (2013). Using "cinéma vérité" (truthful cinema) to facilitate replication and accountability in psychological research. *Frontiers in Psychology*, *4*, Article 872. doi:10.3389/fpsyg.2013.00872

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, Article 160018. doi:10.1038/sdata.2016.18

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Essen, D. C. V., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*, 665–670. doi:10.1038/nmeth.1635